Explainable Facial Emotion Recognition with the use of Vision Transformers

Isidoros Perikos
Computer Engineering
& Informatics
Department
University of Patras,
Computer Technology
Institute and Press
Diophantus
Patras, Greece
perikos@ceid.upatras.gr

Ioannis C. Kollias
Computer Engineer &
Informatics Department
University of Patras
Patras, Greece
st1064886@ceid.upatras.gr

Vaggelis Kapoulas
Computer Technology
Institute and Press
“Diophantus”,
Patras, Greece
kapoulas@cti.gr

Michael Paraskevas
Electrical and Computer
Engineering Department
University of
Peloponnese,
Computer Technology
Institute and Press
“Diophantus”,
Patras, Greece
mparask@cti.gr

***Abstract*—Facial Emotion Recognition (FER) is very important in the field of human-computer interaction and it can greatly help computer systems to interpret and react to human emotions. The analysis of facial expressions and the accurate recognition of their emotional content are highly desired and assistive in a wide spectrum of domains. In this paper, we present a work on the recognition of facial expressions using a hybrid framework that incorporates Vision Transformers (ViT) with Temporal Convolution Networks. The proposed ViT's goal is to extract intricate facial features, whereas the Temporal Convolution Network component effectively captures temporal relationships and aims to enhance the accuracy of facial expression classification. In addition, the LIME technique was used to illustrate the decision-making procedure of the framework utilized. Our framework can achieve an accuracy of 72% on FER2023 dataset, with a strong emphasis on the explanatory power and generalizability of the model.**

Keywords-Facial Emotion Recognition; Vision Transformers (ViT); Explainability; Temporal Convolutional Network (TCN);

I. INTRODUCTION

Emotion recognition from facial expressions forms the backbone of inferences about human intentions and mental states, making it quintessential in human communication [17]. As interactions with machines become increasingly prevalent, teaching computers to perceive human emotions has the potential to revolutionize human-technology interaction. Applications range from healthcare to marketing, where emotionally aware systems can respond actively, leading to more personalized and effective experiences. For instance, Facial Emotion Recognition (FER) in healthcare can assist with early diagnostics in mental health by analyzing even subtle emotional cues, offering the potential to identify conditions like depression and anxiety much earlier than traditional methods. In marketing, FER enables real-time emotional analysis, allowing companies to present tailored product offerings that align with the consumer's emotional state, thereby enhancing the user experience.

Building accurate and reliable FER systems is quite challenging. Emotions are dynamic and change depending on context, which introduces complexity for FER systems [19]. Additionally, the way individuals express emotions can vary significantly based on factors, such as age, gender, ethnicity, and cultural background [20]. External conditions, such as lighting, facial occlusions (e.g., glasses, masks), and head poses further complicate accurate emotion recognition [22]. Moreover, most existing datasets in the literature, while diverse, often fail to account for all these variations, resulting in models that struggle to generalize effectively to real-world scenarios.

Given the sensitivity of the applications, the accuracy and robustness of FER models are paramount. For instance, incorrect emotion detection in healthcare could lead to misdiagnoses, potentially resulting in harmful treatment plans. In fields like customer service or security, undetected or poorly detected emotions can degrade user experiences or even lead to safety issues. As a result, high accuracy, generalization, and reliability are fundamental requirements for trustable FER systems, particularly in critical areas like healthcare, law enforcement, and mental health.

Recent advancements in deep learning and particularly in the formulation of advanced transformer models, have opened new many new possibilities for improving both the accuracy and generalization of FER systems [18]. While initially designed for natural language processing, transformers have demonstrated their potential for image-based tasks by capturing complex patterns and long-range dependencies. Unlike traditional Convolutional Neural Networks (CNNs), which excel at capturing local features, transformers leverage self-attention mechanisms, allowing them to focus on different parts of the face to recognize subtle emotional cues. This makes them well-suited for addressing the nuanced and dynamic nature of emotions. Moreover, the ability of transformers to process sequential and contextual data presents an opportunity to improve FER in environments where emotions fluctuate rapidly [21].

In this paper, we present a work on the recognition of facial expressions using a hybrid framework that incorporates Vision Transformers (ViT) with Temporal

ISBN: 978-1-68558-215-9

Convolution Networks. The proposed ViT's goal is to extract intricate facial features, whereas the Temporal Convolution Network component effectively captures temporal relationships and aims to enhance the accuracy of facial expression classification. In addition, the LIME technique was used to illustrate the decision-making procedure of the framework utilized.

The paper is structured as follows. Section II presents related works. Section III presents the transformer-based approach to analyze facial expressions and recognize their emotional content. After that, Section IV presents the experimental study and the results collected. Section V presents the explainability implementation on the images. Finally, Section V concludes the paper and provides the main directions that future work will explore.

II. Related Works

In recent years, various methods were investigated for FER, and deep learning models, especially recently developed Vision Transformers, show great promise because they allow for the modeling of complex patterns in facial expressions [15][16].

In the work of Chaudhari et al. [1] on the ViTFER, the authors implemented the face emotion recognition system with the help of a vision transformer. The hybrid dataset used by the authors comprised three datasets: FER2013, AffectNet, and CK+48, and is referred to as AVFER. This model has been a fine-tuned Vision Transformer combined with ResNet-18, which was compared. In order to balance samples between classes, data augmentation was performed. In particular, the Vision Transformer (ViT) models proposed in this work with Sharpness-Aware Minimizer reached as high an accuracy as 53.10% in this hybrid dataset and outperform ResNet-18.

VGGNet-based Convolutional Neural Network architecture implemented by Khaireddin et al. [2] achieved best results were realized with a fine-tuning of the hyperparameters and the various techniques of optimization. Using this model, trained on FER2013, gave an accuracy of 73.28% on FER2013, ranking it among the best single-network results. It contains a total of 35,887 grayscale images in size of 48x48, labeled under seven classes of emotions.

Another approach combines Deep CNN with Haar Cascade to capture facial features [3]. This hybrid model, designed for real-time emotion classification, applied pre-processing and data augmentation techniques to optimize training on FER2013, achieved 70% accuracy with significantly reduced training time of 2098.8s.

For ViT-based models, a ViT-CNN hybrid model demonstrated the strength of merging local feature extraction from CNNs with global attention mechanism of ViTs [4]. This model outperformed traditional CNNs by effectively capturing both local and global features, achieving 72.1% accuracy on FER2013.

A more recent contribution is by Wang et al. [5], who, in their submission to the ABAW4 competition, propose an ensemble deep model, CNN-Transformer for facial affect recognition. By construction, the model combines strengths from CNNs that perform well in local spatial feature extraction and strengths from Transformers that capture long-range dependencies and global contextual information. Their method outperformed others with well-balanced performance in a wide range of tasks of facial affect recognition. Their validation set experimental results indicated that their model significantly outperformed the baseline, with a higher F1 score of 0.618 on the LSD task; therefore, this verified the effectiveness of the hybrid design.

Li et al. [6] developed a more powerful hybrid model that combined CNN with a Vision Transformer in facial expression recognition. Their framework leverages the CNN part to extract multiscale local features while the Vision Transformer extracts global relations using the attention mechanism. Besides, they devised a feature integration method and a patch-dropping strategy to further enhance its efficiency and improve the accuracy of recognition. This approach increased the performance significantly in the FER2013 dataset, with an accuracy of about 71.8%, outperforming most models purely based on CNNs or Transformers.

Wang et al. [7] focused on the Vision Transformers with attention mechanisms to further improve the accuracy of emotion recognition. The work underlined the capabilities of transformers to model global context and relationships in facial images, which thus improved results in emotion classification on the FER2013 dataset. Their proposed model attained an accuracy of 74.3%, indicating the effect of an attention-based architecture to capture variations in the facial features.

Kollias et al. [8] present the comprehensive survey on the recognition of face behavior in the wild and analyze different models with respect to operation in unconstrained conditions. Their work has targeted FER challenges in natural conditions and gives a gradual transformation of interest towards transformer-based models and hybrid networks to minimize these challenges, specifically for lesser-controlled facial expression scenarios.

Zhang et al. [9] proposed a real-time face emotion recognition system that combined the attention-based CNN and Transformer components. It was optimized to perform well in real time without a loss in accuracy in detecting fast-changing emotions. Tested on FER2013, it yielded 74.6% accuracy. Thus, hybrid models have also been established for time-critical tasks to be practical.

Zhang et al. [10] developed the attention dual graph convolutional network, which takes facial landmarks as nodes in a graph and models the relationships between them. This is a novel approach, which allows for more ordered representation of the facial features, and because of this reason, it allows for better performance on FER2013 with 90.1%. In fact, their graph-based representations combined with attention mechanisms bring high performance to the model dealing with FER tasks.

Khan et al. [11] proposed the model architecture that hybridized both: the ViT and CNN models. In their model, CNN was used for efficient feature extraction, while with the use of ViTs the possibility of modeling long-range dependencies in facial images can be brought about underneath. The system achieved a recognition accuracy of 76.8% on FER2013, illustrating the viability of hybrid models applied to emotion recognition tasks.

III. Methodology

In this section, we present the approach used for facial emotion recognition and detail the underlying architecture of the hybrid ViT and TCN model, which forms the backbone of this study. The design aims to improve FER accuracy by combining powerful feature extraction techniques with temporal modeling capabilities. Additionally, the LIME technique is employed to enhance the model's explainability, providing insights into its decision-making process.

A. Hybrid framework

We propose a fine-tuned version of customized model named ViTCN. It has a balanced architecture, considering both complexity and performance for particular challenges that come associated with face emotion recognition. More importantly, to further strengthen the model, it is based on explanatory technique Local Interpretable Model-agnostic Explanations (LIME) in analyzing and interpreting model prediction with a view to decision-making. To complement the ViT’s spatial feature extraction, a TCN was integrated to capture temporal relationships across sequences of facial features. The TCN architecture consists of eight convolutional layers, with one initial layer followed by seven convolutional layers in a loop. Each layer uses 1D convolutions, a kernel size of 3, and padding of 1 to process the sequential input data.

1) Data Preprocessing and Augmentation

Data preprocessing involved converting the grayscale images from the FER2013 dataset to RGB format and resizing them from 48x48 to 224x224 to match the input size required by the ViT model. Various data augmentation techniques, including random horizontal flipping, random cropping, color jittering, and random rotation, were employed to increase dataset variability and prevent overfitting. The transformed images were then normalized and converted into vectors, which were loaded into PyTorch’s dataloader for training, validation, and testing.

2) Model Architecture

To create this hybrid model required the definition of its two parts, the first part being ViT and the second part being TCN. The model uses a pre-trained Vision Transformer (*ViTModel*) from the Hugging Face model hub [12][13]. The pre-trained ViT base model utilizes patch size 16x16, with an output hidden dimension of 768. The classifier head of the ViT model is replaced with *nn.Identify()*, meaning the ViT is only used for feature extraction (specifically, the *last_hidden_state* is used). The output sequence, after the features from the ViT have been extracted, passes through a Temporal Convolutional Network. The network consists of 8 convolutional layers: the first initial layer and seven convolutional layers arranged in a loop. Each layer of a TCN consists of one-dimensional convolution with a kernel size fixed to 3 and padding fixed to 1, Rectified Linear Unit (ReLU) non-linearity introducing activation function and dropout rate for regularization at 0.3. The output from TCN acts as an input to *F.adaptive_avg_pool1d(x,1)*, which reduces the sequence length to 1; hence the output size becomes batch_size, channels.

A linear layer (*self.fc)* maps the pooled TCN output to the number of emotion classes. The output from ViT, of shape *(last_hidden_state, shape.(batch_size, seq_len, hidden_dim))* is permuted to change dimensions which is necessary for compatibility with the *Conv1d* layers in the TCN.

3) Hyper-parameters

First, the optimization model relies on the Adam optimizer, as this has adaptive learning rates and momentum, hence being quite suitable when dealing with sparse gradients in large Vision Transformers. In this regard, a starting learning rate of 10^{-4} will be selected to give moderate model parameter updates for stable convergence during training. Apart from that, weight decay with a factor of 10^{-4} was used. In essence, this is an L2 regularization, one that penalizes large weights and hence reduces overfitting; this makes the model more generalized on unseen data.

CrossEntropyLoss criterion was used to assess the performance of the model, which was trained with the help of this criterion. This loss function is quite well adapted for multi-class classification tasks like FER, since it computes the divergence between the estimated probability distribution and the true distribution of the target labels. By minimizing this loss, the model gets gradually trained to provide class probabilities that are close to the actual labels of emotions, thus increasing its accuracy.

A *ReduceLROnPlateau* learning rate scheduler was added to the training pipeline, dynamically adjusting the learning rate based on the performance on the validation set. This ‘min’ mode scheduler would reduce the learning rates when there was no significant improvement of the validation loss. The reduction factor used here was 0.5, halving the current best learning are where the model’s progress started showing a plateau. This will prevent early reductions by giving the model tree epochs of stagnating validation loss before the learning rate is adjusted, as specified by the patience parameter set to 3. The minimum threshold for reducing the learning rate was kept at 10^{-6} to avoid excessive reduction that might impede learning.

Early stopping was used to prevent overfitting and improve model efficiency. This approach monitored the validation loss across epochs and stopped training if any improvement was not witnessed during a certain window. Early stopping was configured to be patient for 7 epochs, meaning it would stop training if the validation loss did not improve for 7 consecutive epochs. Early stopping fired by their model, it reset the weights of the model back to the state representing the epoch with the lowest validation loss. The mentioned technique not only avoided overfitting but also economized computational resources by saving a lot of superfluous epochs of training.

IV. Evaluation Study

The FER2013 dataset is used as a benchmark for developing and testing FER models. It contains a large, diverse set of labeled images across seven basic emotions, making it valuable for overcoming challenges arising from individual and cultural variability. Despite its widespread use, models trained on FER2013 often fail to perform robustly in general scenarios, a challenge that necessitates further improvements in model architectures and training approaches.

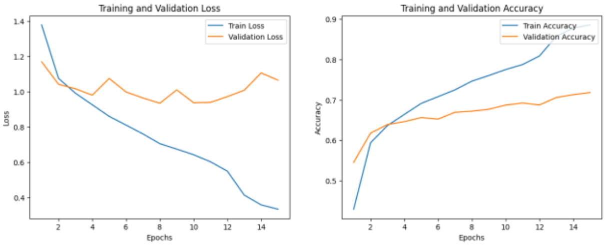

Figure 1. Metric-graphs Results

In Figure 1, the graphs of loss and accuracy for each training and validation epoch are illustrated. We can see that the use of parameters to deal with overfitting was necessary. The evaluation results are quite interesting. We assess the performance of the framework in terms of precision, recall and F1-Score which are reported in Table I. The performance of our model reaches a macro average precision of 74%, with the values of the individual metrics being particularly good.

TABLE I. PERFORMANCE RESULTS

	Classification Report		
	precision	***recall***	***f1-score***
Angry	0.64	0.67	0.65
Disgust	0.89	0.73	0.80
Fear	0.63	0.52	0.57
Happy	0.91	0.89	0.90
Sad	0.57	0.65	0.61
Surprise	0.86	0.79	0.82
Neutral	0.66	0.71	0.68
Macro Avg	0.74	0.71	0.72
Weighted Avg	0.72	0.72	0.72

In addition, the confusion matrix was created, showing the correlations between the actual and predicted labels. The diagonal is identified while it is observed that the fear and sad labels have the highest confusion. Th confusion matrix is illustrated in Figure 2. The confusion matrix provides valuable insight into the performance of the model across different emotion classes. Notably, the model exhibits high accuracy for detecting "Happy" expressions, with a correct prediction rate of 89%, indicating strong performance in recognizing this emotion. However, the matrix also reveals challenges in distinguishing between certain emotions. For instance, the model struggles with differentiating "Fear" and "Sad" expressions, as it correctly identifies "Fear" only 52% of the time, frequently confusing it with "Sad" and "Surprise." Similarly, "Disgust" is often misclassified as "Angry" (18%), suggesting overlap in the facial features associated with these emotions.

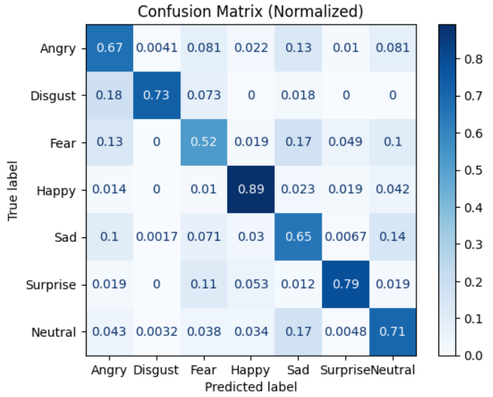

Figure 2. Confusion Matrix

This confusion is indicative of subtle similarities between these emotional expressions that the model may be finding difficult to distinguish. While emotions like "Surprise" and "Neutral" are predicted with relatively high accuracy (79% and 71%, respectively), further improvements could be made for classes like "Fear" and "Disgust." These results suggest that the model would benefit from additional training data, particularly for the underperforming classes, and possibly more refined feature extraction methods to improve its ability to differentiate between similar emotions.

V. EXPLAINABILITY

Explainability in deep learning refers to methods used in an attempt to interpret how models make the decisions. More importantly, for tasks like FER, it is desirable to ascertain that the model decides on the classes of emotions based on meaningful facial features-for example, eyes and mouth. Furthermore, explainability is crucial for building trust in FER systems, as well as for improving the model.

LIME is an explainability technique that provides us insight into model decisions by perturbing parts of the input and observing changes in predictions [14]. This works by creating locally faithful explanations, enabling us to see which regions of the image most drive the model's decision. This is particularly helpful to identify which facial features it relies on to classify emotions-skipping the possibility it learns nonsensical patterns. In this regard, the implementation of the LIME explanation first defines a prediction function, taking a batch of images, preprocesses them by resizing and normalizing, and then passes them through the model in order to compute probabilities for emotion classifications. This prediction function is going to be used by LIME when assessing how model predictions are affected by certain perturbations to the input image.

The key explanation procedure starts by taking an input image and generating slight variations or perturbations of the image. These perturbed images are passed through the model using the prediction function. Subsequently, LIME follows the output response of the model to these variations and selects the most important regions in the image that contributed to the classification decision. It highlights the

Copyright (c) IARIA, 2024. ISBN: 978-1-68558-215-9

boundaries in order to represent areas that had the strongest influence on the prediction. This allow us to interpret visually which parts of the face were of most significance when determining the predicted emotion. Specifically, the explanations generated by LIME illustrate how the model focuses on specific facial regions when classifying emotions These visualizations provide intuitive insights into the model's decision-making process, highlighting the areas of the face that are most influential in predicting each emotion. The yellow-highlighted regions, produced by LIME, confirm that the model is correctly focusing on the most salient facial features typically associated with these emotions, thereby enhancing the interpretability of the model's predictions. The following examples demonstrate the application of LIME in classifying various emotional expressions.

In the example case shown in Figure 3, which illustrates the "Disgust" emotion, the regions of interest are primarily concentrated around the nose and the central part of the face—key areas typically involved in disgust expressions.

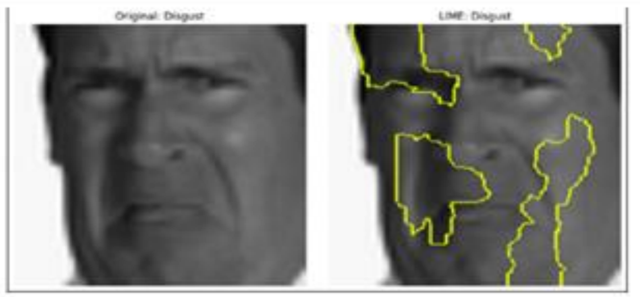

Figure 3. LIME Explanation for Disgust Emotion Prediction.

The LIME output clearly highlights the wrinkling of the nose and the characteristic mouth shape, often frowning or pursed lips. These visual cues further confirm that the model is accurately focusing on the facial features that define disgust. This aligns with human perception, where attention is naturally drawn to the upper face, particularly the nose, when interpreting disgust, as this expression generally centers on the middle of the face. Similarly, in the case for the "Anger" emotion depicted in Figure 4, LIME places emphasis also on the regions around the forehead, eyebrows, and the mouth. As expected, these are the prime areas to consider while one identifies anger, since they have features like furrowed eyebrows and a tense, open mouth both of which are strong indicators of frustration or aggression.

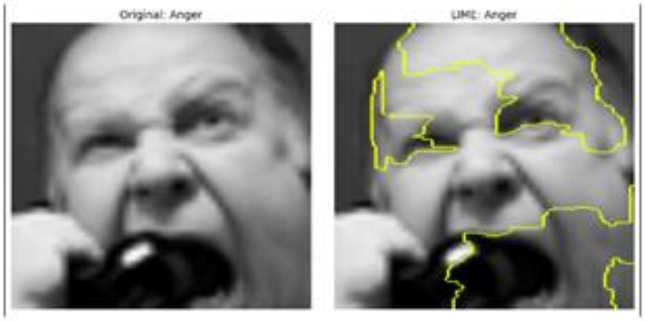

Figure 4. LIME Explanation for Anger Emotion Prediction.

Given the model's attention towards these regions, it convincingly demonstrates that the model has learned to pick out these important features, which adds more credibility to its prediction. Such a visual explanation also agrees with human intuition, since intuitively we associate furrowed eyebrows and tensed facial expressions with anger.

In the example case of the "Sad" emotion in Figure 5, the LIME explanation places significant emphasis on the forehead, eyebrows, and cheeks. The model particularly focuses on key indicators of sadness, such as sagging eyebrows and a drooping mouth. Additionally, the strong attention to the downward gaze suggests that the model is effectively capturing the lowered head posture often associated with sadness. This alignment with typical facial cues reinforces that the model is accurately identifying the relevant features needed for emotion classification.

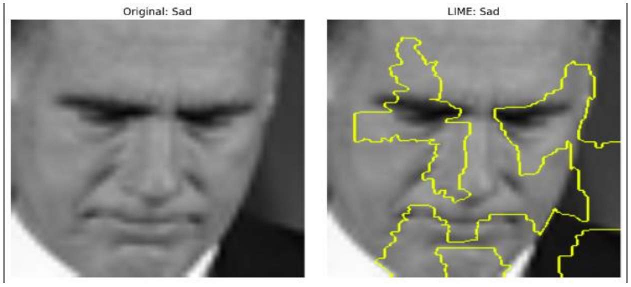

Figure 5. LIME Explanation for Sad Emotion Prediction.

In Figure 6, a case is illustrated for LIME explanation created for a "Surprise" emotion. The LIME explanation has highlighted the forehead, eyes, and mouth regions in this image. Of course, these are important to a surprise expression because wide eyes and a slightly open mouth are common visual keys to this emotion, which the model has learned to focus on appropriately for the expression of surprise. Emphasis on the wide-eyed look and the form of the mouth also aligns with human perception for surprise, further grounding the model's prediction.

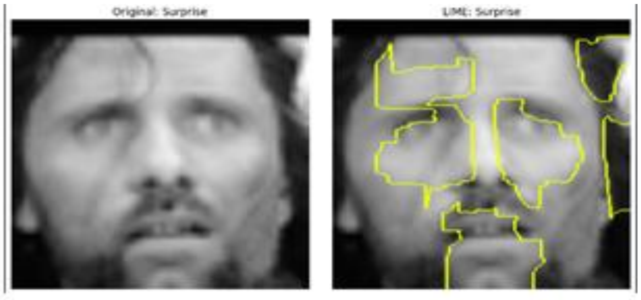

Figure 6. LIME Explanation for Surprise Emotion Prediction.

These LIME visualizations not only show that the model is focusing on relevant and emotion-specific facial features but also offer an interpretable illustration for its predictions. The fact that the LIME highlights align with commonly understood human expressions adds further credibility to the model-that it makes decisions based on appropriate facial cues.

VI. CONCLUSIONS AND FUTURE WORK

Facial expressions form a universal language of emotions, which can instantly express a wide range of emotional states and feelings. The accurate analysis of facial expressions and the precise recognition of their emotional content are highly desired and assistive in a wide spectrum of domains and applications. Although it is natural for humans to interpret facial expressions naturally with little or even no effort, the accurate and robust facial expression recognition by computer systems is still a great challenge. Our work presents a hybrid model for FER by combining both the Vision Transformer and Temporal Convolution Network models. Therefore, this work was successful at achieving high accuracy along with good generalization, Optimization in Vision Transformer was able to extract fine details of a face while embedding TCNs allowed it to identify temporal relationships within the data. In addition, the LIME technique was used to illustrate the decision-making procedure of the framework utilized. Specifically, our framework visualizes explanations with, which explained important facial regions that are responsible for emotions classified within the decision making of the transformer model. The results were very encouraging and indicate that the approach is efficient and accurate in analyzing facial expressions and recognizing their emotional content.

Future research will focus on enhancing generalization by expanding both the size and diversity of the dataset. Additionally, advancing explanation techniques could further increase trust in FER systems, particularly for applications in healthcare and psychological diagnostics. Future directions could involve integrating additional explainability methods, such as SHAP to provide deeper insights into the model's decision-making processes. This combination with LIME, would allow for a clearer understanding of the importance of individual facial regions in each emotion recognition.

ACKNOWLEDGEMENT

This work was partially supported by the Project entitled ”Upgrade and enhancement of the digital services ecosystem of the Panhellenic School Network / Expansion of the myschool system functionalities and completion of the environment for providing advanced digital services to all members of the educational community” with Project Code OPS 5168211

REFERENCES

[1] A. Chaudhari, Y. Khaireddin, and Z. Chen, "Facial emotion recognition: State of the art performance on FER2013". *arXiv Preprint arXiv:2105.03588*, 2021

[2] O.C Oguine, K.J. Oguine, H.I. Bisallah, and D. Ofuani, "Hybrid facial expression recognition (FER2013) model for real-time emotion classification and prediction". *arXiv Preprint arXiv:2206.09509*. 2022

[3] H. Li, M. Sui, F. Zhao, Z. Zha, and F. Wu, "Facial expression recognition using a hybrid ViT-CNN aggregator". In *Springer*, 2021

[4] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, "Visual transformers: Token-based image representation and processing for computer vision", *arXiv Preprint arXiv:2006.03677*, 2020

[5] L. Wang, H. Li, and C. Liu, "Hybrid CNN-Transformer model for facial affect recognition in the ABAW4 challenge". *ArXiv*, 2022

[6] N. Li, Y. Huang, Z. Wang, Z. Fan, X. Li, and Z. Xiao, "Enhanced hybrid vision transformer with multi-scale feature integration and patch dropping for facial expression recognition", *Sensors, 24*(13), 4153, 2024

[7] G. Wang, Y. Zhao, C. Tang, C. Luo, and W. Zeng, "When shift operation meets Vision Transformer: An extremely simple alternative to attention mechanism", *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(2), 2423-2430, 2022

[8] D. Kollias, A. Schulc, and S. Zafeiriou, "Face behavior recognition in the wild: A survey", *IEEE Transactions on Affective Computing, 13*(4), 2244-2263, 2022

[9] J. Zhang, C. Li, G. Liu, M. Min, C. Wang, J. Li, Y. Wang, H. Yan, Z. Zuo, W. Huang, and H. Chen, "A CNN-Transformer hybrid approach for decoding visual neural activity into text", *Computer Methods and Programs in Biomedicine, 214*, 106586. 2021

[10] S. Zhang, Y. Zhang, Y. Zhang, Y. Wang, and Z. Song, "A dual-direction attention mixed feature network for facial expression recognition". *Electronics, 12*(17), 3595,2023

[11] A. Khan, Z. Rauf, A. Sohail, A.R. Khan, H.Asif, A. Asif, and U. Farooq,"A survey of the vision transformers and their CNN-transformer based variants". *Artificial Intelligence Review, 56*, 1-54, 2023

[12] B. Wu, C. Xu, X. Dai, W. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, "Visual Transformers: Token-based image representation and processing for computer vision". *arXiv preprint*, 2020

[13] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database". In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 248-255). IEEE, 2009

[14] E. Hvitfeldt, T.L. Pedersen, and M. Benesty, "*LIME: Local interpretable model-agnostic explanations*", 2022

[15] F.Z. Canal, T.R. Müller, J.C. Matias, G.G. Scotton, A.R. de Sa Junior, E. Pozzebon, and A. Sobieranski, "A survey on facial emotion recognition techniques: A state-of-the-art literature review", *Information Sciences, 582*, 593-617, 2022

[16] Y. Li, J. Wei, Y. Liu, J. Kauttonen, and G. Zhao, "Deep learning for micro-expression recognition: A survey", *IEEE Transactions on Affective Computing*, *13*(4), 2028-2046, 2022

[17] S. C. Leong, Y.M. Tang, C.H. Lai, and C.K.M. Lee, "Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing". *Computer Science Review*, *48*, 100545, 2023

[18] H. Ge, Z. Zhu, Y. Dai, B. Wang, and X. Wu, "Facial expression recognition based on deep learning". *Computer Methods and Programs in Biomedicine, 215*, 106621, 2022

[19] A.V. Savchenko, L.V. Savchenko, and I. Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network". *IEEE Transactions on Affective Computing*, *13*(4), 2132-2143, 2022

[20] S. Kumar, S. Rani, A. Jain, C. Verma, M.S. Raboaca, Z. Illés and B.C. Neagu, "Face spoofing, age, gender and facial expression recognition using advance neural network architecture-based biometric system". *Sensors*, *22*(14), 5160, 2022

[21] L. Xiong, J. Zhang, X. Zheng, and Y. Wang, "Context Transformer and Adaptive Method with Visual Transformer for Robust Facial Expression Recognition", *Applied Sciences*, *14*(4), 1535, 2024

[22] I. Perikos, M. Paraskevas, and I. Hatzilygeroudis, "Facial expression recognition using adaptive neuro-fuzzy inference systems", In *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)* (pp. 1-6). IEEE, 2018